DATA-DRIVEN MODELS FOR UNIAXIAL COMPRESSIVE STRENGTH PREDICTION APPLIED TO UNSEEN DATA

Joaquim Tinoco and António Gomes Correia
C-TAC/Department of Civil Engineering
University of Minho
Campus de Azurém, 4800-058 Guimarãs,
Portugal
E-mail: {jabtinoco|agc}@civil.uminho.pt

Paulo Cortez
Centro Algoritmi/Departamento de Sistemas de Informação
Universidade do Minho
Campus de Azurém, 4800-058 Guimarãs,
Portugal
E-mail: pcortez@dsi.uminho.pt

KEYWORDS
soil improvement, jet grouting, uniaxial compressive strength, data mining techniques, support vector machines

ABSTRACT

Data Mining (DM) techniques have been successfully applied to solve a wide range of real-world problems in different real-world domains, particularly in the field of geotechnical civil engineering. A remarkable example is their use in Jet Grouting (JG) technology. Due to the high number of parameters involved and to the heterogeneity of the soil, JG mechanical properties prediction, as well as columns diameter, are complex tasks. Accordingly, the high learning capabilities of DM, namely of the Support Vector Machine (SVM), were applied in the development of new approaches to accurately perform such tasks. This paper aims to assess the SVM model performance trained to predict Uniaxial Compressive Strength (UCS) of JG samples extracted directly from JG columns, when applied to a new set of records collected from a new JG work not contemplated in the database used during the model learning phase. The achieved results highlight the importance of the model domain applicability, as well as the restrictions and recommendations for its generalization when applied to new JG work data not contemplated in the training dataset.

INTRODUCTION

The capability to automatically learn from data is a very attractive approach to extract useful knowledge. Therefore, in the last decade the use of machine learning has spread rapidly throughout computer science and beyond. Machine learning have shown to be very useful to solve real-world complex problems. These tools, supported on advanced statistics analysis, are usually known as Data Mining (DM) techniques and have been applied successfully in different knowledge domains, such as in web search, spam filters, recommender systems and fraud detection (Domingos 2012). Taking advantage of its strong flexibility to deal with high dimensionality problems, also in geotechnical civil engineering field DM techniques were applied to solve complex problems (Gomes Correia et al. 2013, Miranda et al. 2011).

Following these successful applications, we applied for the first time (from the best of our knowledge) DM tools in the study of Jet Grouting (JG) technology, namely for mechanical properties prediction of both laboratory formulations (Tinoco et al. 2011), and field mixtures (Tinoco et al. 2012b;a). The developed model for Uniaxial Compressive Strength (UCS) prediction of JG mixtures (Tinoco 2012) are the target of the present study, which represent a contribution for a better understanding of JG technology and therefore for its technical and economic efficiency.

In few words, JG technology is classified as a grouting on ground improvement methods and is defined as placement of a pumpable material (normally a cementitious material) directly into the subsoil, without previous excavation (Choi 2005). The cinematic energy of the drilling fluid cut the soil, allowing its mixture with the injected grout. At the end, a new material, also known as *soilcrete*, with a controlled geometry structure is obtained, presenting better physical and mechanical proprieties when compared with natural soil. Indeed, JG is actually a viable and economically attractive solution for a large diversity of geotechnical problems when conventional injection methods are unsuitable, unsafe or too expensive, being nowadays one of the most soft soil improvement methods widely applied worldwide (Poh and Wong 2001).

However, despite of all its advantages and many years of experience, JG design is still a complex and difficult task to perform due to the high number of parameters involved as well as the heterogeneity of the soil. Aiming to overcome this problem, new approaches were proposed based on advanced statistics analysis, usually know as DM techniques, for JG mechanical properties prediction of both laboratory formulations and field mixtures, as well as for JG column diameter (Tinoco 2012).

An important issue related to any data-driven model is its generalization capability, which is measured by the

quality of the predictions for unseen data and is it related with the model's applicability. Accordingly, and although these issues have been taken into account during the learning phase of the models through a cross-validation approach, as detailed in Tinoco (2012), in the present paper we evaluate how the developed model for UCS prediction of *soilcrete* samples behave when applied over a new set of records collected from JG columns related to a JG work not contemplated in the database used during the model learning phase. In this paper, we intended, on one hand, to measure the influence of model domain applicability and, on the other hand its generalization capacity, particularly when applied to data from a JG work different those used in the model training.

MODEL CHARACTERIZATION

In the present work, the performance of the data-driven models proposed in Tinoco (2012) for UCS prediction of *soilcrete* mixtures are here applied over new data collected from JG columns of a new JG work. Particularly, we analyse the one who achieved the most interesting results, either in terms of model performance and interpretability, that is, the one developed using the Support Vector Machine (SVM) algorithms (Tinoco et al. 2014), termed in this paper as *SVM-UCS.Field* model.

A detailed and full description of *SVM-UCS.Field* model as well as a complete characterization of the dataset used to train it can be found in Tinoco (2012). Following are just summarized the main aspects related to this model, as well as to the dataset used to train it.

The input variables considered in *SVM-UCS.Field* model were: JS - Jet System; W/C - Water Cement ratio; ω - water content of the mixture; $\%Clay$ - percentage of clay in the natural soil; t - age of the mixture; $1/\rho_d$ - inverse of dry density of the mixture; C_{iv} - volumetric content of cement; e - void ratio of the mixture; $n/(C_{iv})^d$ - relation between mixture porosity and volumetric content of cement, and its statistics can be found in (Tinoco et al. 2014).

Figure 1 depicts the relationship between experimental versus predicted values by *SVM-UCS.Field* model, which achieved the most interesting result considering the model's accuracy and interpretability. As shown, *SVM-UCS.Field* model was able to predict 81% of the records with an absolute deviation lower than 2 MPa. Moreover, for 13% of the remaining 19% of the records where the absolute deviation is higher than 2 MPa, the predicted value was above the real value. This means that the prediction is on the safety side.

Taken into account that model accuracy is not enough for its full assessment, Figure 2 shows and compares the relative importance of each input variable according to *SVM-UCS.Field* model, which was based on a global sensitivity analysis (Cortez and Embrechts 2011). Interpreting this figure, we can see that the $n/(C_{iv})^d$, JS, t and $\%Clay$ are the four key variables in UCS prediction of *soilcrete* mixtures (Tinoco et al. 2014). Among them it is identified one that is related to the soil type (%Clay), another related with the JG process (JS) and two others related to the JG mixture, namely its age and the relation $n/(C_{iv})^d$ that combines the porosity and cement content effect. In other words, to predict UCS of *soilcrete* mixtures, the model ask for information about the soil to be improved, how the improvement was performed and the actual conditions of the obtained mixture. Moreover, it is also interesting to observe that such variable ranking has a physical explanation and is empirically acceptable.

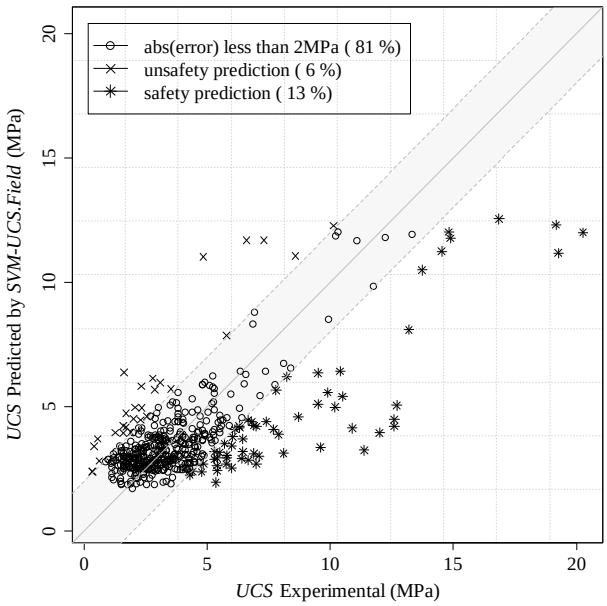

Figure 1: Relationship between UCS experimental versus predicted values by *SVM-UCS.Field* model

SVM MODEL PERFORMANCE OVER UNSEEN DATA

As summarized in previous section, *SVM-UCS.Field* model achieved the best performance in UCS prediction of *soilcrete* mixtures, when applied to a dataset with 472 rows collected from five different JG works, for which was applied a 20-fold cross-validation approach during its learning process (Tinoco 2012).

Meanwhile, new records about UCS of *soilcrete* mixtures from another JG work (further identified by letter C) were gathered and compiled into a new dataset with 31 rows. These new records were used in the present study to assess *SVM-UCS.Field* model generalization capacity when applied to data from a new JG work not contemplated in the dataset used for its training and measure the influence of model domain applicability. Table 1 summarises the main statistics of this new dataset over which *SVM-UCS.Field* model was ap-

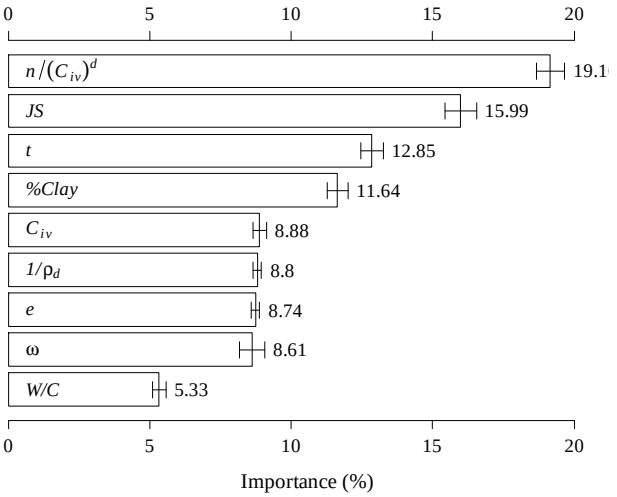

Figure 2: Relative importance of each input variable according to *SVM-UCS.Field* model, quantified by 1-D sensitivity analysis (Tinoco et al. 2014)

plied to predict UCS. Comparing this table with those in (Tinoco et al. 2014) that summarizes the statistics of *SVM-UCS.Field* model attributes we can see that the range of $n/(C_{iv})^d$, $1/\rho$ and e variables of the new dataset is not totally inside of the range of the dataset used during the learning process of the models. Particularly, the first one is a key variable in the model and was based on it that some records were defined as outside model domain. This particular situation will allow us to assess *SVM-UCS.Field* model performance when applied to different/new conditions (extrapolation) when compared to the model learning conditions. In other words, will allow us to measure the influence of model domain applicability. Accordingly, it was defined that a new record is inside model domain if only the value of all input variables are within model attributes range. If at least one input variable is out of the model attributes range, then this record is considered as outside model domain. We tested other strategies for the definition of inside/outside model domain samples, such as clustering (e.g. k-nearest neighbour algorithm), but achieved identical results and hence we adopt the simpler domain range rule.

Moreover, in order to establish more evidence-based conclusions about the generalization capacity of SVM algorithm, as well as SVM models domain applicability influence, an external cross-validation, using the JG work site as a criterion, was applied for all six JG works actually available. This mean that we iteratively remove one work site from the database, train the model with five JG works and apply it in the prediction of the removed work site records. In these cases, t and $\%Clay$ were the main variables that defined some records as outside model domain.

Figure 3 depicts the relationship between UCS experimental versus SVM models predicted values for all records inside model domain, under an external cross-validation approach as above defined.

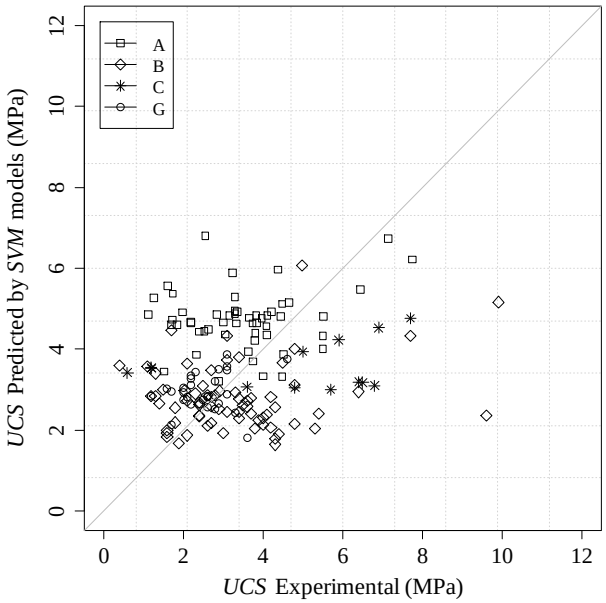

Figure 3: Relationship between UCS experimental versus predicted values by SVM models for records inside model domain

The achieved results show that SVM models have some difficulties to predict accurately UCS of *soilcrete* samples collected from JG columns from a new JG work not contemplated in the model learning phase. The dispersion shown in Figure 3 was also observed for those records defined as outside models domain.

Therefore, these results lead us to underline that SVM models trained to predict UCS of *soilcrete* mixtures should be careful applied in UCS prediction of new records related to a JG work not contemplated in the model learning phase. This results underline the current practice in JG works, where before built the project columns, some test columns need to be built in order to define each JG parameter that accomplish project requirements. It should be stressed that these test columns are compulsory for each new JG work.

In order to corroborate this statement, an additional experience was performed. This experiment consists on applying an external cross-validation approach as above describe, but for the model learning phase additional to the records related to the five JG works, it was also added a percentage of records from the sixth work. Accordingly, and taken into account the amount of data normally collected from JG test columns, a percentage of 15% and 25% were considered. Moreover it was also performed an experiment with a training dataset containing 75% of the records from the sixth work. For example, and following the strategy of 25%, the SVM model is trained with all records from JG works A, B,

Table 1: Summary statistics of both input and output variables of the new records (from JG work C) to predict UCS of *soilcrete* mixtures

Variable	Minimum	Maximum	Mean	Standard Deviation
JS	2.00	2.00	2.00	0.00
W/C	0.91	0.91	0.91	0.00
ω	4.00	67.00	23.97	18.21
$\%Clay$	40.00	40.00	40.00	0.00
t	32.00	85.00	58.16	20.49
$1/\rho_d$	4.15E^{-4}	1.15E^{-3}	6.33E^{-4}	1.97E^{-4}
C_{iv}	0.15	0.15	0.15	0.00
e	0.13	2.08	0.72	0.54
$n/(C_{iv})^d$	12.14	72.94	40.33	16.72
UCS	0.60	10.20	5.24	2.07

C, D and E plus 25% of the records of JG work F. Then, this model is tested in the remaining 75% records of JG work F.

Table 2 compares the achieved performance considering 0%, 15%, 25% and 75% of the records from the sixth work in the training dataset, using Mean Absolute Deviation (MAD), Root Mean Square Error (RMSE) and Coefficient of Correlation (R^2) (Tinoco 2012) as a error measure. From Table 2 results, we can see that, as expected, as more data from the new JG work were used in the model training, better is its performance in the prediction of new records related to this new JG work. As shown, adding just 15% of records from the new JG work, we can improve significantly the model performance. Thus, in order to ensure a better reliability when the model is applied to data from a JG work not contemplated during its learning phase, it is recommended to retrain the model adding some data from the new JG work to the training dataset, corresponding to the data obtained with the JG test columns. As a balance, between the trade-off of needing more data and having a good prediction performance, we believe that the 15% strategy is quite useful. Moreover, it is also observed that SVM models have a great capability to improve UCS prediction of soilcrete mixtures along the construction works as data will be available. It was shown that when around 75% of the records related to the new JG work are available, SVM models are able to predict with high accuracy UCS of *soilcrete* mixtures for the remaining columns. In this conditions it was achieved the performance depicted in Figure 4, for which correspond an $R^2 = 0.63$.

CONCLUSIONS

Data Mining (DM) techniques are being considered as a good alternative to the traditional statistics to solve complex problems. Indeed, even in the geotechnical civil engineering field these tools have been successfully applied. One particular application is its use in Jet Grouting (JG) technology, characterized not only by a great versatility, but also by its complexity due to the high number of parameters involved and to soil heterogeneity. Related to this issue, some data-driven models were developed to predict Uniaxial Compressive Strength (UCS) of *soilcrete* samples over time, which achieved extraordinary results.

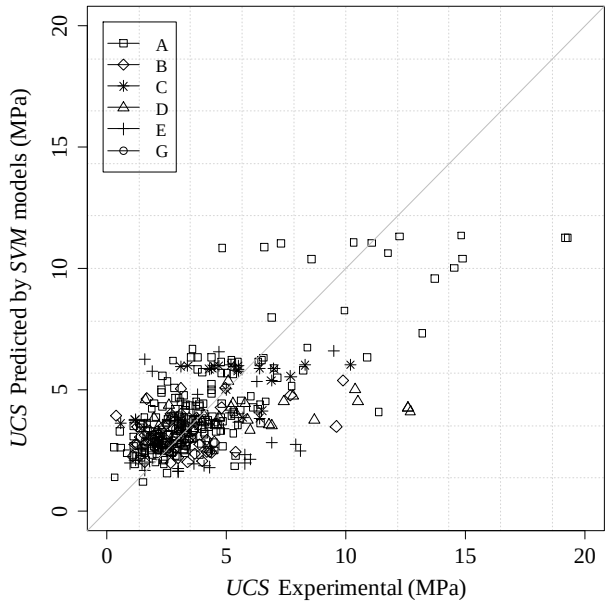

Figure 4: Relationship between UCS experimental versus predicted values by SVM models under an strategy of 75%

Supported on these models, particularly on the SVM, this work underlined two main conclusions when a given model is applied in the prediction of new results from a new construction site. The first one is the importance

Table 2: SVM models performance under 0%, 15%, 25% and 75% strategies

Metric	0%	15%	25%	75%
R^2	0.21	0.45	0.46	0.63
MAD (MPa)	1.58	1.53	1.44	1.34
$RMSE$ (MPa)	2.51	2.14	2.04	1.79

of the model domain applicability that can lead us to reach wrong conclusions if such factor is not considered when performing predictions on unseen data. The second one is that the accomplishment of the model domain applicability may not be sufficient for an accurate prediction of unseen data, particularly if these new records are from a new JG work not contemplated during the model learning. It was shown that in order to ensure a better generalization performance, it is fundamental to include some data (e.g. 15%) related to new JG work (for which it is intended to make predictions), in the dataset used for model training. This observation underlines the need of creating some test columns for each new JG work before starting the works, which is a compulsory practice in JG works. Thus, before applying a given data-driven model, namely SVM, it is strongly recommended to carefully analyse its domain applicability. Moreover, when applying it to new records related to a JG work not contemplated during the model learning phase, it is also advised to retrain the model with some data from this new JG work (data from JG test columns), allowing it to learn the particular characteristics of this new JG work. Additionally, it was also observed that SVM models have a great capability to improve UCS prediction of *soilcrete* mixtures along the construction works as data will be available.

ACKNOWLEDGEMENTS

The authors wish to thank to "Fundação para a Ciência e a Tecnologia" (FCT) for the financial support under the Pos-Doc grant of strategic project PEst-OE/ECI/UI4047/2011. Also, the authors would like to thank the interest Tecnasol-FGE company for providing all data needed.

REFERENCES

Choi R., 2005. *Review of the Jet Grouting Method.* Bachelor dissertation, Faculty of Engineering and Surveying, University of Southern Queensland, Australia.

Cortez P. and Embrechts M., 2011. *Opening Black Box Data Mining Models Using Sensitivity Analysis.* In *2011 IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2011.* IEEE, Paris, France, 341–348.

Domingos P., 2012. *A few useful things to know about machine learning. Communications of the ACM*, 55, no. 10, 78–87.

Gomes Correia A.; Cortez P.; Tinoco J.; and Marques R., 2013. *Artificial Intelligence Applications in Transportation Geotechnics. Geotechnical and Geological Engineering*, 31, no. 3, 861–879.

Miranda T.; Comes Correia A.; Santos M.; Sousa L.; and Cortez P., 2011. *New Models for Strength and Deformability Parameter Calculation in Rock Masses Using Data-Mining Techniques. International Journal of Geomechanics*, 11, 44–58.

Poh T. and Wong I., 2001. *A Field Trial of Jet-grouting in Marine Clay. Canadian Geotechnical Journal*, 38, no. 2, 338–348.

Tinoco J., 2012. *Application of Data Mining Techniques to Jet Grouting Columns Design.* Ph.D. thesis, School of Engineering, University of Minho, Guimarães, Portugal.

Tinoco J.; Gomes Correia A.; and Cortez P., 2011. *Application of Data Mining Techniques in the Estimation of the Uniaxial Compressive Strength of Jet Grouting Columns over Time. Construction and Building Materials*, 25, no. 3, 1257–1262.

Tinoco J.; Gomes Correia A.; and Cortez P., 2012a. *Jet Grouting Deformability Modulus Prediction using Data Mining Tools.* In M. et al. (Ed.), *Proceedings of the 2nd International Conference on Transportation Geotechnics.* Taylor & Francis Group, Hokkaido, Japan, 168–173.

Tinoco J.; Gomes Correia A.; and Cortez P., 2012b. *Jet Grouting Mechanicals Properties Prediction using Data Mining Techniques.* In *Proceedings of the 4th International Conference on Grouting and Deep Mixing.* ASCE, New Orleans, Louisiana, USA, 2082–2091.

Tinoco J.; Gomes Correia A.; and Cortez P., 2014. *Support Vector Machines Applied to Uniaxial Compressive Strength Prediction of Jet Grouting Columns. Computers and Geotechnics*, 55, 132–140.